Medusa at the University of Illinois at Urbana-Champaign: A Digital Preservation Service Based on PREMIS

Kyle Rimkus
University of Illinois at Urbana-Champaign
44 Library
Urbana, IL 61801
(+1) (217) 300-3842
rimkus@illinois.edu

Thomas Habing
University of Illinois at Urbana-Champaign
155 Grainger Engineering Library
Urbana, IL 61801
(+1) (217) 244-4425
thabing@illinois.edu

ABSTRACT

The Medusa digital preservation service at the University of Illinois at Urbana-Champaign provides a storage environment for digital content selected for long-term retention by content managers and producers affiliated with the Library in order to ensure its enduring access and use. This paper reports on Medusa development, with emphasis on the research processes that informed key decisions related to its design, the central role of PREMIS metadata in its architecture, and future directions of integrating PREMIS management into a Fedora repository architecture. In so doing, it describes a strategy of digital preservation content management that draws strength from the creation and management of comprehensive PREMIS preservation metadata records.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: Systems issues

Keywords

Digital Preservation, Medusa, PREMIS

1. Digital Preservation Research at UIUC

1.1 National Digital Information Infrastructure Preservation Program (NDIIPP) Grants

In 2004, digital preservation was still a young library practice. The Open Archival Information System specification was relatively new, having been finalized in 2002 [1], while the working group to establish Preservation Metadata: Implementation Strategies, or PREMIS, had been formed in 2003 [5]. With such developments catalyzing interest and a sense of urgency in the library community, the National Digital Information Infrastructure and Preservation Program, or NDIIPP [9] arose in the United States to support efforts to develop the tools and methodologies necessary to initiate and sustain digital preservation activities in libraries and other historical memory institutions across the country.

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.

The University of Illinois at Urbana-Champaign received two rounds of NDIIPP funding from 2004-2007 and 2007-2010, respectively. These grants, undertaken in partnership with the Graduate School of Library and Information Science, the National Center for Supercomputer Applications, OCLC Inc., peer institutions, and several state libraries throughout the United States, were known together as the ECHO DEPository or "ECHODep" [6], and featured research on web archiving tools, repository software interoperability, the semantics of preservation metadata, and tools for file evaluation and assessment. In particular, grant activities related to the development of the Hub and Spoke, or "HandS" suite of tools for repository interoperability [8] established a broad knowledge-base of digital preservation concepts, technologies, and standards among institutional library faculty and staff. Lessons learned from this project went on to inform current digital preservation repository development, so they bear some further explanation.

1.2 Hub and Spoke, or HandS Framework Tool Suite

Repository research in the first phase of ECHODep focused on evaluating the interoperability characteristics of five platforms then available to digital library practitioners: the open source DSpace, Fedora, Greenstone, and ePrints platforms, as well as the hosted OCLC Digital Archive service. Taking it as a given that the packaging and transfer of digital content and its associated metadata from one repository platform to another has become an inevitable event in the lifecycles of contemporary digital materials, and that this poses definite digital preservation risks, the research team undertook a variety of experiments moving sets of digital objects from one platform to another, thereby gaining knowledge of the metadata and file packaging profiles for ingest and export native to each respective system.

These activities continued in Phase Two with the development of a suite of tools for moving content between repository platforms. Given the name of "Hub and Spoke" or "HandS," the Illinois team developed an open-source, Java-based tool with a graphical user interface for packaging data as it moved from a central "hub" to any number of "spokes," or ingest destinations. This approach was taken to mitigate the implicit inefficiency of Phase One's method:

"To reduce the complexity of interoperability, the Hub and Spoke uses a common packaging format for interchange of digital resources between different repositories. Digital packages coming from a repository are transformed into this common format before any further processing, and digital packages are transformed from the common format into the native repository format when being

placed into a repository. The idea is to reduce an N^2 problem into a 2N problem…" [4].

The hub through which all transfers of content passed, however, was not a repository itself, but a Metadata Encoding and Transmission Standard, or METS file that grew every time an object made the trip from one spoke repository to another, by retaining all original and crosswalked metadata files produced during transfers and organizing them into a structured XML document. This approach was distinguished by several key factors:

1. the reliance on PREMIS for digital preservation metadata;
2. the reliance on MODS for descriptive metadata;
3. the packaging of PREMIS, MODS, and other associated metadata and file information in METS wrappers, to describe relationships between files and significant events in their respective lifecycles; and
4. the implicit assertion that digital object packages designed to serve a long-term preservation need ought to be repository-independent.

Three METS profiles that embody this approach were registered with the Library of Congress [12]. This work also contributed to the publication of *Guidelines for using PREMIS with METS for exchange* [3], and went on to define in large part the approach taken to building Medusa, with one important exception. Namely, Medusa's developers decided to implement their service without METS in favor of a purely PREMIS-centric model for managing digital preservation metadata. This will be discussed in some detail following a broad overview of Medusa.

2. Digital Preservation Services at UIUC

2.1 The Need

By 2010, the Library still had much in common with the problems investigated by ECHODep's Hub and Spoke research – that is, it was (and remains) an organization with a variety of content producers and systems for managing and delivering content to patrons. As of October, 2012, a preliminary assessment conducted by the Library's Preservation Unit identified at least eight platforms in the library for managing access to locally produced digital library content, including ContentDM, DSpace, ARTStor, the Internet Archive, the HathiTrust Digital Library, Archon, and Olive ActivePaper.

In addition, the master files created for the access systems listed above were not all being stewarded in a single, unified storage environment. While the majority of them existed on file servers managed by the Library's Information Technology group, a significant number of them could be found on optical media in an off-site storage facility, on hard drives on staff members' shelves, or on file servers leased from external companies (a full analysis of the preservation needs of these collections prior to Medusa ingest is being prepared for future publication). Given these circumstances, significant motivation has existed for several years within the Library to introduce a more reliable infrastructure for the long-term preservation of locally managed digital content.

After six years of research into the theory and practice of digital preservation, the University of Illinois Library nevertheless found itself without a full-fledged service to ensure enduring access to its own digital collections. It relied on several external preservation services for specific formats of digital collections, and has, in the intervening years, introduced others. Taken together, these include the HathiTrust for digitized books, the Web Archiving Service for web content, and the LOCKSS Alliance and Portico for subscription electronic journals. Medusa fills a special need not addressed by these external services. Namely, its collections consist of unique digital content that the University of Illinois Library has the need, authority, and responsibility to store, manage, and preserve locally. This includes but is not limited to text, image, audio, and video preservation master files created for Library units by its digital reformatting operations, as well as electronic records acquired by its repositories of special collections.

2.2 Medusa is Born

At the heart of Medusa is a Collection Registry written in Ruby on Rails (public collection entries may be viewed at https://medusa.library.illinois.edu/). This web application allows preservation managers to survey digital collections outside of the repository, assess their potential risk factors, and ingest them into a secure preservation storage environment. The Open Archival Information System (OAIS) reference model, as well as the complementary Trustworthy Digital Repository certification framework, have both strongly informed Medusa's infrastructure. Its designated community of users consists primarily of managers of digital content to which the University of Illinois Library allows access. This includes curators of archival and special collections, subject specialists, and managers of repositories of scholarly content, as well as internal units that produce and manage digital files.

3. PREMIS in Medusa

3.1 The PREMIS Philosophy

The philosophy underpinning Medusa is that all of its objects are self-describing. Medusa objects and the events that occur to them throughout their respective lifecycles are represented entirely by PREMIS records. By virtue of their rigorously maintained metadata, Medusa digital objects are therefore able to stand on their own independently of their underlying repository software and hardware.

To achieve this, every digital asset stewarded in Medusa – whether a content or metadata file – is assigned a unique ID and an associated PREMIS file. In contrast, for example, to the common repository practice of storing digital content files on a file server and metadata in a database, the "self-contained object" approach favors managing objects in their native formats using a common storage layer. Risk increases when the constituent parts of digital objects are split up across a variety of systems subject to their own specific threats, a problem that Medusa's architects seek to avoid.

For similar reasons, Medusa's technical team has also decided to forego METS. While many practitioners deploy PREMIS within METS wrappers [2][13], leveraging METS's richness in expressing relationships between files with PREMIS's affordances in file and representation-level preservation metadata, others have pointed out significant challenges in bundling PREMIS in METS, due to the lack of shared and agreed-upon practices for managing the complexity of hierarchical item and metadata records, as well as redundancies between the two schemas [7].

With respect to tag overlap, metadata managers who implement PREMIS in METS must choose whether to place file information such as file type and file size, among other things, in METS tags,

PREMIS tags, or duplicate them in both. The overlap in available tags is so significant, in fact, that one could argue that the only thing that significantly differentiates METS from PREMIS is the lone required METS tag-set called the "structMap," which provides a structural map of METS packages and describes the relationships between the metadata and bit-streams they contain.

Furthermore, expressing complex relationships in consistently valid METS XML is challenging. The nesting of multiple XML files within a single METS wrapper requires complex validation procedures, the maintenance and debugging of which cause bottlenecks in production workflows, not to mention significant overhead in programming reliable METS tools tailored to local needs.

In other words, there's a lot to keep track of when utilizing METS as the centerpiece of a digital preservation system. PREMIS, on the other hand, offers a flatter data structure which nevertheless enables similarly rich and complex modeling of relationships between digital objects. Namely, its "relationship" tag, by allowing a "relationshipType" and a "relationshipSubType" to refine it, presents the ability to link any one object to any other object with elegant simplicity. What one ends up with, in effect, is the ability to define any imaginable relationship between any single entity and any number of others. Similar flexibility is also enabled, via the same mechanism, to defining Events, Agents, and Rights specific to assets in one's digital preservation environment. (For a closer look at PREMIS controlled vocabularies in Medusa, visit the Medusa wiki at https://wiki.cites.uiuc.edu/wiki/display/LibraryDigitalPreservation/PREMIS+Controlled+Vocabularies). Inspired by current trends in linked data and its underlying concepts, the Medusa implementation team has committed itself to utilizing the PREMIS "relationship" tag to define structured relationships between objects and their constituent parts, and sees it as preferable to the strictly hierarchical METS alternative.

3.2 PREMIS Implementation

Two simplified figures illustrate how Medusa uses PREMIS to model the relationships within a digital object. Figure 1 displays a PREMIS collection-level "representation" record with a root-level MODS metadata record describing the collection itself, as well as a Rights statement and Event metadata testifying to the circumstances of the collection's instantiation in Medusa, in this case its ingest by the "Agent" Tom Habing. Figure 2 shows a related object, namely, an Archival Information Package that belongs to the same collection. This object's PREMIS metadata shows that the object consists of the front and back of a digitized postcard with both archival JPEG2000 and access JPEG image files available for both, and a MODS XML descriptive metadata record. The object is also linked to data declaring its Rights information. The archival images link to an Event detailing their ingest, or capture, into Medusa.

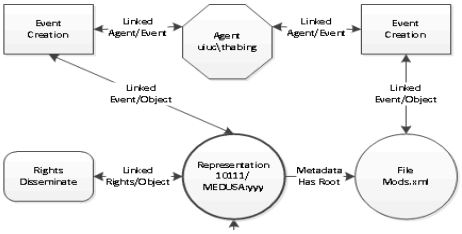

Figure 1. PREMIS Model for a Collection Described by a Single MODS File

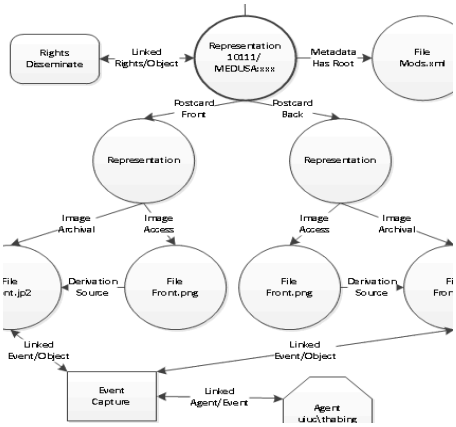

Figure 2. Simplified PREMIS Model Representing the Front and Back a Digitized Postcard

4. Fedora Implemenation

Considerable work has also gone into how best to map Medusa's PREMIS-based object models into a broader Fedora architecture for its object-level preservation store. The approach under consideration is often referred to in Fedora circles as the 'atomistic' model, in which each Fedora object contains no more than a single datastream (not including the standard datastreams such as "DC" or "RELS_EXT"). In this model, each PREMIS entity will reside in a single Fedora object; for example, each PREMIS Agent will be modeled as a Fedora object containing a PREMIS Agent XML file as its only datastream. Likewise, each PREMIS Event will be modeled as a Fedora object containing a single PREMIS Event XML datastream. PREMIS Representation Objects are modeled similarly. The only exceptions to this are PREMIS File Objects. File Objects contain two datastreams, the PREMIS File Object XML file and the actual content file itself. Figure 3 is a simple illustration of this. The boxes are Fedora objects with their PIDs shown at the top and their datastreams shown below. The RELS_EXT relationship between objects are shown with arrows. Note that the relationships between PREMIS objects are converted into RELS_EXT predicates in of the form *relationshipType.subType* in a custom namespace.

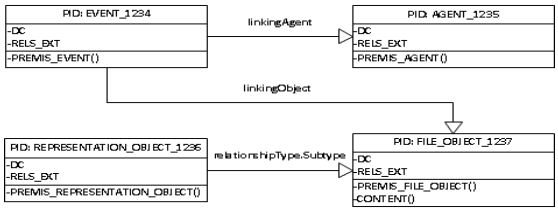

Figure 3. PREMIS Event linked to Agent as Fedora objects

To avoid redundantly encoding the same data in multiple places, all relationships between PREMIS entities will be encoded using Fedora RELS_EXT relationships. PREMIS-based submission packages (SIPs) into the repository may use PREMIS linking mechanisms, but as part of the ingest process these relationships will be encoded in RELS_EXT streams and removed from the corresponding PREMIS entities. Conversely, the repository will provide archival information packages (AIPs) where these relations are represented using PREMIS linkages derived from their respective Fedora RELS_EXT, in keeping with the guiding principle of repository-independent file packages.

The directionality of RELS_EXT relationships is driven by the desire to minimize the frequency of modifications to any individual object over time. For example, it is anticipated that

objects may accumulate many events throughout the course of their existence in Medusa. When an event occurs to a repository object, a new PREMIS Event record will be created as a Fedora object. This event will have a RELS_EXT link back to its associated object, but the RELS_EXT for the object itself will not be modified to point to all of its events. However, all events associated with a given objects can still be discovered by querying the repository's underlying RDF triple store.

5. Current Status

As of April 2013, the Medusa development team has implemented a functioning collection registry, bit-level ingest feature, and an object-level PREMIS packager. All project code is available in a Github repository at https://github.com/medusa-project. Development is ongoing, following an agile methodology with representatives of the Library's Preservation Unit providing user specifications to programmers who work in weekly sprints to add and enhance features. Although much administrative information is limited to authenticated system users, interested researchers and library practitioners may view unrestricted collection registry records as https://medusa.library.illinois.edu/. In addition, the project team maintains repository policies and specifications in a regularly updated wiki at https://wiki.cites.uiuc.edu/wiki/display/LibraryDigitalPreservation/Medusa+Digital+Preservation+Service.

Currently, Medusa managers are drafting Submission, Archival, and Dissemination Information Package specifications with content producers to inform the next steps of object-level content management in Medusa. One important looming decision is that of broader repository architecture. While all progress to date has been made developing local code to meet preservation managers' immediate needs, the Library's Software Development Group is, as mentioned above, planning next steps to include a Fedora digital object management layer. More specifically, the Medusa development team is considering the Hydra application stack of Fedora, SOLR, Blacklight, and Ruby on Rails [10][11] as a viable framework for broader repository services. Hydra, so named because of its flexibility to support multiple "heads," or customizable services on top of a single Fedora repository, is seen as a potential long-term solution to the challenge described throughout this paper – that is, the proliferation of access repository services at Illinois without a central hub in which to exercise institutional control over the life-cycle of digital library objects.

6. Conclusion

In conclusion, the establishment of a digital preservation service at the University of Illinois at Urbana-Champaign Library is firmly rooted in lessons learned from six years of NDIIPP grant research from 2004-2010. This research underscored the importance to reliable digital preservation management of PREMIS metadata and the practice of packaging digital objects in a repository-independent manner. This philosophy has carried over to the intellectual and technical modeling of Medusa, the Library's developing digital preservation service. Medusa relies heavily on a comprehensive implementation of PREMIS. It currently consists of a web-accessible collection registry that allows preservation managers to, among other things, ingest packages of files into a long-term storage environment. Medusa development is ongoing, with the goal of providing a model for fully integrating PREMIS metadata into the life cycle of digital content stewarded for long-term access at the University of Illinois at Urbana-Champaign Library.

7. REFERENCES

[1] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)." CCSDS Secretariat, Jan-2002.

[2] R. S. Guenther, "Battle of the Buzzwords," *D-Lib Magazine*, vol. 14, no. 7/8, Jul. 2008.

[3] R. S. Guenther, R. Wolfe, O. Brandt, M. Enders, T. G. Habing, F. Lazzarino, C. Redding, and J. Riley, "Guidelines for using PREMIS with METS for exchange," Library of Congress, Washington, DC.

[4] T. Habing, J. Eke, M. A. Cordial, W. Ingram, and R. Manaster, "Developments in digital preservation at the University of Illinois: The Hub and Spoke architecture for supporting repository interoperability and emerging preservation standards," *Library Trends*, vol. 57, no. 3, pp. 556–579, 2009.

[5] PREMIS Editorial Committee, "PREMIS Data Dictionary for Preservation Metadata: version 2.2," Jul. 2012.

[6] J. Unsworth and et al, "ECHO DEPository Technical Architecture Phase 1 Final Report: Report of project activities from Fall 2004 through 2007," University of Illinois at Urbana-Champaign in partnership with OCLC, Dec. 2008.

[7] S. Vermaaten, "A Checklist and a Case for Documenting PREMIS-METS Decisions in a METS Profile," *D-Lib Magazine*, vol. 16, no. 9/10, Sep. 2010.

[8] "UIUC Echodep Hub and Spoke Documentation." [Online]. Available: http://dli.grainger.uiuc.edu/echodep/hands/index.html. [Accessed: 25-Sep-2012].

[9] "Digital Preservation (Library of Congress)." [Online]. Available: http://www.digitalpreservation.gov/index.php. [Accessed: 25-Sep-2012].

[10] "Medusa - Hydra - DuraSpace Wiki." [Online]. Available: https://wiki.duraspace.org/display/hydra/Medusa. [Accessed: 13-Nov-2012].

[11] "The Hydra Project - Hydra - DuraSpace Wiki." [Online]. Available: https://wiki.duraspace.org/display/hydra/The+Hydra+Project. [Accessed: 13-Nov-2012].

[12] "ECHO Dep Generic METS Profile for Preservation and Digital Repository Interoperability, ECHO Dep METS Profile for Web Site Captures, and ECHO Dep METS Profile for Master METS Documents." [Online]. Available: http://www.loc.gov/standards/mets/profiles/00000016.html. [Accessed: 17-Jan-2013].

[13] "PREMIS in METS Toolbox." [Online]. Available: http://pim.fcla.edu/. [Accessed: 31-Jan-2013].